DEIM Forum 2011 B5-6

ニュースにおけるバーストキーワードの話題への集約

高橋　佑介†　　宇津呂武仁††　吉岡　真治†††

† 群馬工業高等専門学校専攻科生産システム工学専攻　〒 371-8530 群馬県前橋市鳥羽町 580 番地
†† 筑波大学大学院システム情報工学研究科　〒 305-8573 茨城県つくば市天王台 1-1-1
††† 北海道大学大学院 情報科学研究科　〒 060-0808 北海道札幌市北区北 8 条西 5 丁目

あらまし　ウェブ上では情報爆発が起こり，ニュース記事に限っても，日々，多種多様な話題が発信され続けている．このような多様な情報から特に重要な情報を選びだす手法として Kleinberg のバースト解析アルゴリズムを用いることができ，これにより，ある時期において，特に頻繁に報道された話題に関する記事や，頻出したキーワードを抽出することができる．本論文では，Kleinberg のバースト解析アルゴリズムによって得られたバーストキーワード集合に対してさらに，複数のバーストキーワードが共通に出現するニュース記事のクラスタリングを行うことにより，バーストキーワードを話題ごとのまとまりに集約した．これによって，各キーワードのバーストの背景となった話題，および，各バーストキーワード間の関係の把握が容易になった．また，ここで，バーストキーワード群を利用して，ニュース記事のクラスタリングを行う手法を導入した．その結果，既存手法よりも少ない次元数の特徴ベクトルによって，既存手法と同程度のクラスタリング性能が達成できることがわかった．
キーワード　ニュース, バースト, トピック, 集約

Aggregating Bursty Keywords in News Stream into Topics

Yusuke TAKAHASHI†, Takehito UTSURO††, and Masaharu YOSHIOKA†††

† Advanced Production Systems Engineering Course, Gunma National College of Technology, Maebashi, 371-8530, Japan
†† Grad. Sch. of Systems and Information Engineering, University of Tsukuba,　Tsukuba, 305-8573, Japan
††† Graduate School of Information Scienceand Technology, Hokkaido University,　Sapporo, 060-0808, Japan

Abstract　Among various types of recent information explosion, that in news stream is also a kind of serious problems. Especially, when one wants to detect a kind of topics that are paid much more attention than usual, it is usually necessary for him/her to carefully watch every article in news stream at every moment. In such a situation, it is well known in the field of time series analysis that Kleinberg's modeling of "bursts" is quite effective in detecting burst of keywords. Based on this argument, this thesis proposes a technique of aggregating such bursty keywords in news stream according to the topic that is closely related to each bursty keyword. The proposed technique is based on clustering news articles which share a group of bursty keywords. This approach enables to easily identify the event occurring in the real world which caused the burst of keywords. It also helps to understand the relation among several bursty keywords. We compare the proposed technique of clustering news articles based on bursty keywords with a baseline clustering technique, and show that we can achieve comparative clustering performance even with keyword vectors of a much smaller number of dimension.
Key words　news, burst, topic, aggregation

1. はじめに

ウェブ上の世界を始めとして，膨大な情報が溢れ，いわゆる情報爆発が起こっている．文献 [3], [7] でも述べられているように，この現状の課題を克服するような取り組みがなされている．ウェブ上のニュース記事に限っても，同様に多くの情報が流れているため，そこで例えば一週間のニュースの中でも大きな話題の動きを簡潔にまとめて提示できるようなシステムがあれば，ニュース記事を閲覧するにあたり大きな時間の節約になるだろう．本論文は，このように，情報爆発における情報の集約という背景に基づいて研究を行った結果を報告するものである．

ネット上の文書ストリームでは，時間を問わず文書情報が送

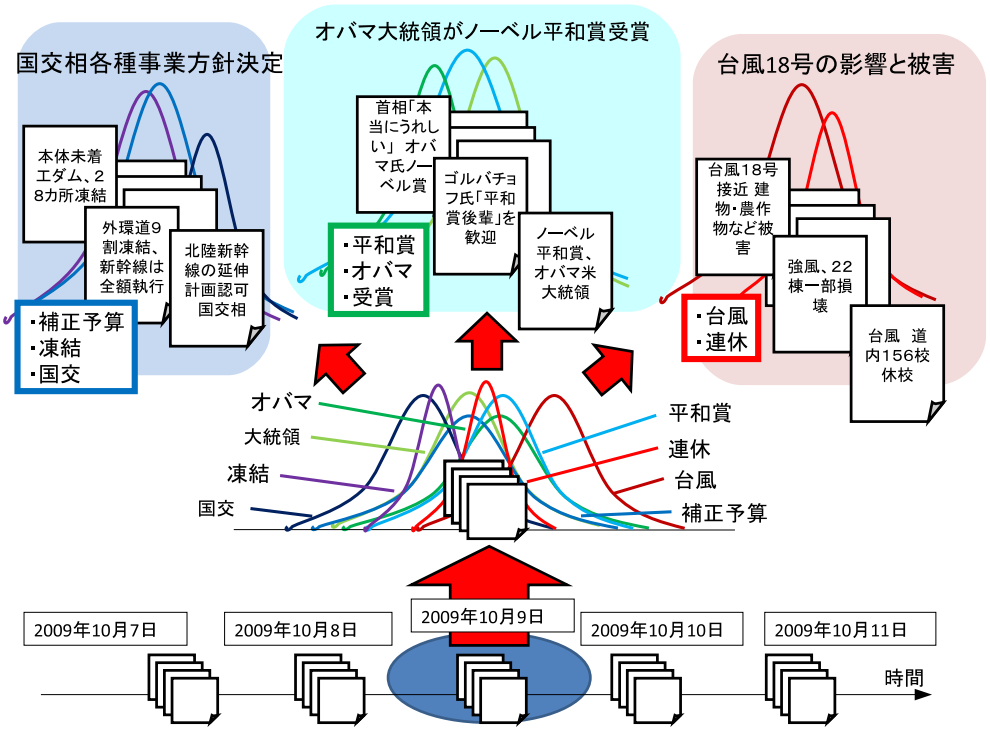

図 1　バーストキーワードの話題への集約

られている．そこでは，ある時からある話題に関する記述が急激に増加するような現象が起こることがあり，こういった現象を，ある話題に関するバーストと呼ぶ [4]．例えば，世間で騒がれるような出来事が起こると，その時期には，その話題に関連したニュース記事が爆発的に多くなり，ニュース記事ストリームにおいてその話題がバーストする．したがって，ニュース記事におけるバーストの時期を解析できれば，そういった世の中の異変を掴むのに役立つと考えられる．

Kleinberg のバースト解析 [4] を用いることによりそのような時系列解析が可能となり，さらにその応用として，ある日における各キーワードのバースト度を求めてランク付けすることができる [9]．これにより，例えば各日におけるバースト度の高い上位数件のキーワード（バーストキーワード）を得ることができる．ここから，得られた結果を時系列に沿って並べることで，キーワードの時系列推移を見ることができるはずである．

そこで，得られた上位数件のバーストキーワードを調べると，異なるキーワードであっても，それらは同じ話題を示すものが多いことがわかった．このことより，キーワードのバーストは，そのキーワードを関連概念とした話題のバーストが背景にあると考えた．キーワードの集約化ができれば，より俯瞰的に，話題そのもののバーストを捉えることができるようになるはずである．そして，時系列推移を見たとき，単なるキーワードの推移ではなく，話題の推移が見られるようになると考えた．本論文ではこの考えを元に，各日ごとにバーストキーワードの出現したニュース記事をクラスタリングし，それを用いてバーストキーワードを話題ごとに集約する手法を提案する．提案手法の枠組みは図 1 のようになる．

以下，2. では，Kleinberg のバースト解析 [4] について説明する．まず，バースト解析の手法の一つである enumerating バーストについて説明し，そこからバーストランクを定義する．3. では，バーストキーワードの集約について，その手順と評価結果について述べる．バーストキーワードの集約は，バーストキーワードを含む記事をクラスタリングすることによって行う．その際，文書間類似度を，提案手法と比較手法の二種類定義して，記事のクラスタリング結果を評価して比較する．4. では，関連研究について述べる．最後に 5. では，まとめと今後の課題を述べる．

2.　Kleinberg のバースト解析

本節では，Jon Kleinberg の考案した，バースト解析アルゴリズム [4] について述べる．

2.1　バースト解析とは

バースト解析は，典型的には，電子メールやネット上のニュース記事などの文書ストリームに対して適用されるアルゴリズムである．文書ストリームを時系列に沿って観測すると，ある期間において，あるキーワードを含む文書の時間軸方向の密度が高くなるような状態が発生する．このような状態をバーストという．Kleinberg のバースト解析アルゴリズムを用いれば，文書ストリーム中のあるキーワードのバースト期間と非バースト期間とを自動で切り分けることが可能になる．図 2 に文書ストリームにおけるバーストのイメージ図を示す．

2.2　enumerating バースト

enumerating バーストのアルゴリズムは，離散時間で送られる文書のまとまり (文書バッチ) に対して適用される．つまり，常に連続的に，単体で文書が送られる状況ではなく，ある期間をおいてまとまった文書が送られてくるような状況において適

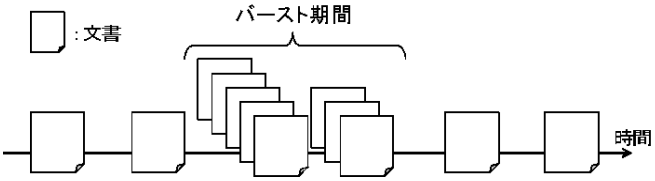

図 2　バーストのイメージ

用される．

最も簡単なモデルでは 2 状態オートマトン $\mathcal{A}^2$ を定義し，2 つの状態を非バースト状態 q_0，バースト状態 q_1 とおく．入力に対して状態が遷移することにより，2 つの状態を切り分ける仕組みである．目的とする文書(注1)を「関連文書」，そうでない文書を「非関連文書」として扱い，バーストか否かは，文書バッチ中の関連文書の割合によって決まる．

解析期間において，n 個の文書バッチ B_1〜B_n が離散時間で送られてくる状況を考える．t 番目のバッチを B_t とし，そのバッチに含まれる文書の数を d_t とおく．文書バッチには関連文書と非関連文書が含まれ，B_t に含まれる関連文書の数を r_t とおく．解析期間における全ての文書の数 D は $D = \sum_{t=1}^{n} d_t$，解析期間における全ての関連文書の数 R を $R = \sum_{t=1}^{n} r_t$ と表すことができる．

次に，オートマトンの 2 状態にそれぞれ期待値を割り当てる．初期状態である非バースト状態 q_0 には，分析期間全体を見たときの期待値 $p_0 = R/D$ を割り当てる．バースト状態 q_1 には，p_0 にパラメータ s をかけた値である $p_1 = p_0 s$ を割り当てる．ただし，$s > 1$ であり，$p_1 \leqq 1$ となるような s でなくてはならない．s の値が小さいほど，バッチ中の関連文書の割合が低くてもバーストと見なされやすくなる．

解析は，n 個のバッチが与えられたときの，状態の系列を通るためのコスト計算によって行う．考えられる状態の系列のうち，最も系列のコストが小さいものが解となり，その系列の状態に応じて，バースト期間と非バースト期間を決定する．

状態遷移は d_t と r_t が入力となって決まる．状態の系列は $q = (q_{i_1}, \ldots, q_{i_n})$ と表され，q_{i_n} は，n 番目のバッチによって決定された状態 q_i $(i = 0, 1)$ である．バッチ中の関連文書が二項分布 $B(d_t, p_i)$ にしたがって現れるという考えに基づき，状態 q_i にいることに対してコストを与える関数 $\sigma(i, r_i, d_t)$ を以下のように定義する．

$$\sigma(i, r_t, d_t) = -\ln\Big[\binom{d_t}{r_t} p_i^{r_t} (1 - p_i)^{d_t - r_t} \Big]$$

この関数は，入力 r_t と d_t，及び状態 q_i $(i = 0, 1)$ によってコストが決まる．$p_1 > p_0$ であり，t 番目のバッチ中における関連文書の出現確率 r_t/d_t が p_1 に近ければ $\sigma(1, r_t, d_t) < \sigma(0, r_t, d_t)$ となり，コストの低いバースト状態 q_1 が選ばれる．逆に，r_t/d_t が p_0 に近ければ $\sigma(1, r_t, d_t) > \sigma(0, r_t, d_t)$ となり，非バースト状態 q_0 が選ばれる．

ただし，閾値付近の入力が続くなどして頻繁に状態遷移が起こると，途切れ途切れにバースト状態と非バースト状態が切り替わり不自然である．そのために，状態遷移を妨げるための関数 $\tau(i, j)$ を定義する．

$$\tau(i, j) = \begin{cases} (j - i)\gamma & (j > i) \\ 0 & (j \leqq i) \end{cases}$$

τ は，パラメータ γ によって調節されるが，特に理由がない場合は $\gamma = 1$ とする．

以上に述べた，ある状態 q にいることに対してコストを与える関数 σ と，状態遷移にペナルティを課す関数 τ を使って，状態の系列 q を通るためのコスト関数を定義する．

$$\begin{aligned} c(\, q \,|\, r_t, d_t\,) &= \Big(\sum_{t=0}^{n-1} \tau(i_t, i_{t+1}) \Big) + \Big(\sum_{t=1}^{n} \sigma(i_t, r_t, d_t) \Big) \\ &= \Big(\sum_{t=0}^{n-1} \tau(i_t, i_{t+1}) \Big) + \\ &\quad \Big(\sum_{t=1}^{n} -\ln\Big[\binom{d_t}{r_t} p_{i_t}^{r_t} (1 - p_{i_t})^{d_t - r_t} \Big] \Big) \end{aligned}$$

オートマトン $\mathcal{A}^2$ は二つのパラメータ s，γ によって決まることから，$\mathcal{A}^2_{s,\gamma}$ と表記される．本実験では，$s = 2$，$\gamma = 1$ として $\mathcal{A}^2_{2,1}$ のオートマトンを用いている．

2.3　バーストランク

バーストランクとは，各キーワードのバーストの強さを計算し，その大きさによってキーワードをランク付けしたものである．

まず，各キーワードのバーストの強さを測るものとしてバースト度を定義する．期間 t_1〜t_2 におけるバースト度は以下の式で表される．

$$\sum_{t=t_1}^{t_2} \big(\sigma(0, r_t, d_t) - \sigma(1, r_t, d_t) \big)$$

つまり，期間 t_1〜t_2 に，非バースト状態 p_0 だけを通る系列のコスト $\sum_{t=t_1}^{t_2} \sigma(0, r_t, d_t)$ と，バースト状態 p_1 だけを通る系列のコスト $\sum_{t=t_1}^{t_2} \sigma(1, r_t, d_t)$ との差によって求められる．

以上のようにしてキーワードごとに求めたバースト度についてキーワードを昇順にソートすることで，期間 t_1〜t_2 における各キーワードのバーストランクを作成することができる．本実験では，1 日ごとにバーストランクを算出している．

3.　バーストキーワードの集約

本節では，ある日に求められたバーストキーワードを話題ごとに集約するための手順と，それを評価した結果について述べる．

3.1　ニュース記事のクラスタリングによるバーストキーワードの集約手順

バーストキーワードの集約は，以下の手順で行った．

(注1)：例えば，特定のキーワードを含む文書．

（1） ある日において，バーストランク上位 20 個のバーストキーワードを含むニュース記事を収集する．

（2） 人手で記事を選別する．

（3） ニュース記事間の類似度を測定し，この類似度を用いて記事をクラスタリングする．

（4） クラスタごとにバーストキーワード集合を作成し，話題ごとに集約されたバーストキーワード集合とする．

上記 4 つの手順の詳細を以下に示す．

まず，キーワードの選定においては，ニュース記事本文に対して茶筌 (注2) および IPAdic (注3) を用いて形態素解析を行った後，名詞と動詞を抽出した．さらに，それと並行して，Wikipedia エントリ名を抽出した．ただし，1 文字だけのもの，数字のみのもの，記号だけのもの，および，形態素解析誤りに起因する文字列断片は除外した．

次に，評価対象とする日付を 10 日分選定し，日付ごとにバーストランクの高い上位 20 個のバーストキーワードを取得する．ただしこの際,「6 月」や「8 月」など，月を表すキーワードが頻繁にバーストランクに上位に現れたため，これらはノイズとなると考え，バーストキーワードからは除外した．

なお，本論文では，2009 年 6 月 1 日 ∼2010 年 5 月 31 日の 1 年間の期間において収集したニュース記事集合(注4)から選定したニュース記事を評価対象として用いる．また，ニュース記事集合に対する df(文書頻度)，idf(逆文書頻度) の統計量を算出する際には，この記事集合の全ニュース記事を用いる．

次に，各日において，バーストキーワードと密接に関連する記事を人手で選別するため，はじめに，バーストキーワードを 2 つ以上含む記事を選出する．その後，システムの出力したクラスタリング結果の評価のため，まず人手でニュース記事を話題単位でクラスタリングして，参照用正解クラスタを作成する．そして，それを用いて，各日における主要な話題との関連性が低い記事を除外する．具体的には，参照用正解クラスタのうち，主要な話題のクラスタに含まれなかった記事が除外される．以上の手順により，評価実験に用いるニュース記事集合および参照用正解クラスタを用意する．評価実験の対象とした日付および記事数を表 1 に示す．

表 1　評価対象の日付およびニュース記事数

日付	記事数	日付	記事数
2009 年 6 月 1 日	49	2009 年 12 月 19 日	29
2009 年 8 月 1 日	54	2010 年 1 月 13 日	126
2009 年 9 月 1 日	112	2010 年 3 月 1 日	101
2009 年 10 月 9 日	59	2010 年 4 月 12 日	67
2009 年 11 月 1 日	24	2010 年 5 月 1 日	47

次に，ボトムアップクラスタリングによって，選別された記事集合を話題ごとにクラスタリングする．各クラスタの中心は，各クラスタに含まれる記事の特徴ベクトルの重心によって求めた．また，クラスタの分割数を決める際には，クラスタリング結果に対する評価尺度 (F 値) が最大となる分割数を人手で選んだ．

ここで，記事間の類似度としては，提案手法，および，比較手法の 2 通りを評価した．提案手法においては，20 個のバーストキーワードの有無を二値で表現した 20 次元ベクトルを記事の特徴ベクトルとして定義し，特徴ベクトル間の内積，および，余弦によって類似度を測定した．比較手法においては，記事中に出現したキーワードを特徴ベクトルの次元とし，各次元のキーワードの tf-idf 値を各次元の値として特徴ベクトルを定義し，特徴ベクトル間の内積，および，余弦によって類似度を測定した．比較手法における特徴ベクトルの次元数は，平均して 20 以上となる．

最後に，クラスタリングの結果を用いて，バーストキーワードを集約する．この時点で，記事は話題ごとにクラスタリングされている．したがって，クラスタごとに，記事に含まれるバーストキーワードを収集してバーストキーワード集合を作成することにより，話題ごとにバーストキーワードが集約された結果が得られる．

3.2 ニュース記事のクラスタリング結果の評価

ニュース記事のクラスタリング結果の評価においては，参照用正解クラスタを用いる．評価尺度には，以下に定義する，再現率，適合率，F 値を用いる．

$$\text{再現率} = \frac{\text{出力された各クラスタに含まれる記事組のうち，参照用正解クラスタに含まれる記事組数の和}}{\text{各参照用正解クラスタに含まれる記事組数の和}}$$

$$\text{適合率} = \frac{\text{出力された各クラスタに含まれる記事組のうち，参照用正解クラスタに含まれる記事組数の和}}{\text{各システム出力クラスタに含まれる記事組数の和}}$$

$$F\text{ 値} = \frac{2}{\frac{1}{\text{再現率}} + \frac{1}{\text{適合率}}}$$

両手法の性能を比較するため，実験で評価した 10 日分のマクロ平均とミクロ平均を算出した．その結果を表 2 に示す．

表 2　両手法のクラスタリング性能の比較 (10 日分の平均 (%))

		提案手法		比較手法	
		内積	余弦	内積	余弦
マクロ平均	再現率	97.1	94.7	88.5	88.4
	適合率	86.6	91.6	90.5	87.7
	F 値	91.0	**92.3**	88.0	87.5
ミクロ平均	再現率	95.3	91.0	85.4	84.2
	適合率	77.7	83.7	86.6	76.5
	F 値	85.6	**87.2**	86.0	80.2

この結果より，提案手法によって特徴ベクトルを作成し，余弦によって記事間類似度を測ることによってクラスタリングする手法が最も F 値が高いことが分かる．提案手法では特徴ベ

(注2)：http://chasen-legacy.sourceforge.jp/

(注3)：http://sourceforge.jp/projects/ipadic/

(注4)：日経新聞 (http://www.nikkei.com/)，朝日新聞 (http://www.asahi.com/)，読売新聞 (http://www.yomiuri.co.jp/) の各新聞社のサイトから収集した 56,503 記事，38,758 記事，および，62,684 記事の合計 157,945 記事．

クトルの次元数が少ないにも関わらず，両手法のクラスタリング性能は比較手法と同等以上である．これより，バーストキーワードのみを参照するだけでも，各記事の話題をある程度とらえることができているといえる．また，提案手法においては，余弦によって類似度計算を行う方がクラスタリング性能が高いが，比較手法においては，内積を用いる方が良いことがわかった．

3.3 バーストキーワードの集約結果

ニュース記事のクラスタリングによって，バーストキーワードの集約をすることができた．類似度として余弦を用いた場合について，提案手法によってニュース記事をクラスタリングした結果，および，バーストキーワード集約結果を，表 3，および，表 4 に示す．この結果から分かるように，人手によって作成した参照用正解クラスタの各々が，それぞれ固有のバーストキーワードを多く含む場合には，クラスタリング性能が高くなり，逆に，複数の参照用正解クラスタ間で，多くのバーストキーワードが共有されている場合には，クラスタリング性能が低くなっている．

なお，2009 年 9 月 1 日における「衆議院選挙」と「民主党新政権発足に向けて」という 2 つの話題や，2010 年 4 月 12 日における「バンコク騒乱」と「バンコク騒乱で邦人男性死亡」などの，因果関係を持つ 2 つの話題を，それぞれ「衆議院選挙」や「バンコク騒乱」といった，1 つの大きな話題としてみなした参照用正解クラスタを作成した場合には，クラスタリングの性能がより高くみなせることが分かる．しかし，今回の場合，このような 2 つの話題に関する記事が，それぞれ同程度の記事数で分布していたため分割して扱うことにした．

4. 関 連 研 究

Kleinberg のバースト解析を利用せずにバーストキーワードを集約する関連研究がある．文献 [1] では，キーワードに対して生成確率分布を仮定してバーストを判定し，確率分布の形状が類似しているキーワード同士をまとめることで，バーストキーワードの抽出とイベント単位の集約を行う方法を提案している．また，文献 [6] では，Twitter(注5)のツイートログからバーストキーワードを抽出し，共起性の高いバーストキーワード同士を集約することにより，Twitter 上のトレンドを提示する枠組みを提案している．一方，文献 [2] では Kleinberg のバースト解析を用いているが，文書クラスタリングが主眼である．ここでは，特徴ベクトルの次元として，分析期間中にバーストしたキーワードを用い，重みとしては，tf-idf 値などに加え，分析期間中におけるキーワードのバースト度を付加する手法を提案している．その他，バーストを多角的な視野でとらえることを主目的として，異なる言語間のニュース記事ストリームのバーストパターンを比較し，相関性のあるバーストトピックを抽出する手法 [8] なども提案されている．

(注5)：http://twitter.com/

5. おわりに

本稿では，バーストキーワードを含むニュース記事をクラスタリングし，その結果を用いてバーストキーワードを話題ごとに集約する手法を提案した．本手法では，クラスタリングがニュース記事単位で行われているため，キーワードのバーストの背景にある話題ごとに，キーワードを集約することができる．また，クラスタリングされたニュース記事を閲覧することにより，どのような話題に基づいてキーワードが集約されたかを容易に知ることが可能である．

さらに，記事のクラスタリングを行う際，バーストキーワードのみを用いて記事間の類似度を測定する手法を評価した．評価実験の結果，記事中の全てのキーワードを次元として tf-idf の重みを用いて測定した類似度と比較して，次元数がより少なくても同程度のクラスタリング性能を維持できることが分かった．

今後は，Wikipedia を知識源として，記事中の観点の時系列分布を分析する手法 [5] を導入することにより，バーストキーワードおよびニュース記事群の集約性能をより高める方式を開発する予定である．また，現在は最適なクラスタ数の決定を手動で行なっているため，その自動方式について検討を行う．

文　献

[1] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu. Parameter free bursty events detection in text streams. In *Proc. 31st VLDB*, pp. 181–192, 2005.

[2] Q. He, K. Chang, E-P. Lim, and J. Zhang. Bursty feature representation for clustering text streams. In *Proc. 7th SDM*, pp. 491–496, 2007.

[3] 喜連川優. 特定領域研究「情報爆発（Info-plosion)」: 本格稼働から 2 年を経過して. 情報処理, Vol. 49, No. 8, pp. 881–888, 2008.

[4] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proc. 8th SIGKDD*, pp. 91–101, 2002.

[5] 牧田健作, 横本大輔, 宇津呂武仁, 福原知宏. トピックに関する話題の時系列分布に着目したブログ分析. 第 3 回データ工学と情報マネジメントに関するフォーラム—DEIM フォーラム— 論文集, 2011.

[6] M. Mathioudakis and N. Koudas. TwitterMonitor: Trend detection over the Twitter stream. In *Proc. SIGMOD*, pp. 1155–1157, 2010.

[7] 鳥澤健太郎, 中川裕志, 黒橋禎夫, 乾健太郎, 吉岡真治, 藤井敦, 喜連川優. キーワードサーチを超える情報爆発サーチ — 自然言語処理で価値ある未知をマイニング —. 情報処理, Vol. 49, No. 8, pp. 890–896, 2008.

[8] X. Wang, CX. Zhai, and R. Sproat X. Hu. Mining correlated bursty topic patterns from coordinated text streams. In *Proc. 13th SIGKDD*, pp. 784–793, 2007.

[9] 吉岡真治. 多言語ニュースの対象分析のための Wikipedia 活用手法の研究. 第 23 回人工知能学会全国大会論文集, 2009.

表 3　10 日分のニュース記事のクラスタリング結果及びバーストキーワード集約結果 (1)

日付	クラスタ番号	集約されたキーワード (キーワードが出現した参照用正解クラスタの ID)	参照用正解クラスタとの対応 (参照用正解クラスタの話題及び ID)	ニュース記事のクラスタリング結果の評価 (再現率/適合率/F 値 (%))
2009 年 6 月 1 日	1	申請／民事／米政府／再生／ゼネラル・モーターズ／民事再生法／ gm ／破産／破綻／連邦／適用／条／連邦破産法 11 条 (ID:1)	GM 破綻 (ID:1)	100/100/100
	2	新型インフルエンザ／補正／新型／感染／インフルエンザ (ID:2)	新型インフルエンザ流行 (ID:2)	
2009 年 8 月 1 日	1	上田 (ID:1,2,5)，公認 (ID:2,5)，京都 (ID:3,4,5)，青森 (ID:4,5)，自民党／訴える／衆院選／衆院／訴え／麻生首相／県連／自民／立候補 (ID:5)	衆院選に向けての動き (ID:5)	96.9/95.3/96.1
	2	メートル (ID:1,2,3,4)，選手権／決勝 (ID:1,2,4)，上田 (ID:1,2,5)，水泳 (ID:2)，公認 (ID:2,5)	リコー全英女子オープン (ID:1) ／世界水泳選手権 (ID:2) ／全国高校野球選手権大会 (ID:4)	
	3	選手権／決勝 (ID:1,2,4)，京都 (ID:3,4,5)，甲子園 (ID:4)，青森 (ID:4,5)	全国高校野球選手権大会 (ID:4)	
	4	選手権／決勝 (ID:1,2,4)，上田 (ID:1,2,5)，ゴルフ (ID:1,3)	リコー全英女子オープン (ID:1)	
	5	メートル (ID:1,2,3,4)，ゴルフ (ID:1,3)，京都 (ID:3,4,5)	近畿南部豪雨被害 (ID:3)	
2009 年 9 月 1 日	1	民主／民主党 (ID:1,2,3)，自民党／衆院選／政権／麻生／自民／衆院／選挙／行政 (ID:1,2,3,5)，当選／復活／比例 (ID:2,3)，政権交代／交代 (ID:2,3,5)，学期 (ID:3,4)	消費者庁発足 (ID:1) ／民主党新政権発足に向けて (ID:2) ／衆議院議員総選挙 (ID:3) ／自民党総裁選 (ID:5)	69.8/63.2/66.4
	2	民主／民主党 (ID:1,2,3)，新型インフルエンザ／インフルエンザ／新型 (ID:1,2,3,4)，自民党／衆院選／政権／麻生／衆院／選挙 (ID:1,2,3,5)，政権交代／交代 (ID:2,3,5)	民主党新政権発足に向けて (ID:2) ／衆議院議員総選挙 (ID:3) ／自民党総裁選 (ID:5)	
	3	民主／民主党 (ID:1,2,3)，新型インフルエンザ／インフルエンザ／新型 (ID:1,2,3,4)，政権／麻生／自民 (ID:1,2,3,5)	消費者庁発足 (ID:1) ／民主党新政権発足に向けて (ID:2)	
	4	民主／民主党 (ID:1,2,3)，自民党／政権／麻生／自民／選挙／行政 (ID:1,2,3,5)	消費者庁発足 (ID:1) ／民主党新政権発足に向けて (ID:2) ／衆議院議員総選挙 (ID:3)	
	5	民主 (ID:1,2,3)，政権／自民 (ID:1,2,3,5)，政権交代／政権 (ID:2,3,5)	民主党新政権発足に向けて (ID:2)	
	6	自民党／麻生／自民／選挙／行政 (ID:1,2,3,5)	自民党総裁選 (ID:5)	
	7	新型インフルエンザ／インフルエンザ／新型 (ID:1,2,3,4)，学期 (ID:3,4)	新型インフルエンザ流行 (ID:4)	
	8	衆院 (ID:1,2,3,5)，民主党 (ID:1,2,3)，復活 (ID:2,3)	民主党新政権発足に向けて (ID:2)	
	9	行政 (ID:1,2,3,5)，民主党 (ID:1,2,3)	民主党新政権発足に向けて (ID:2)	
2009 年 10 月 9 日	1	連休 (ID:1,3)，オバマ大統領／核兵器／平和／ノーベル平和賞／ノーベル／受賞／大統領／ノーベル賞／オバマ (ID:3)	オバマ大統領がノーベル平和賞受賞 (ID:3)	100/100/100
	2	国土，凍結 (ID:1,2)，補正予算／今年度／予算／国交／来年度／補正／前原 (ID:2)	国交相各種事業方針決定 (ID:2)	
	3	台風 (ID:1)，国土／凍結 (ID:1,2)，連休 (ID:1,3)	台風 18 号の影響と被害 (ID:1)	

表 4　10 日分のニュース記事のクラスタリング結果及びバーストキーワード集約結果 (2)

日付	クラスタ番号	集約されたキーワード (キーワードが出現した参照用正解クラスタの ID)	参照用正解クラスタとの対応 (参照用正解クラスタの話題及び ID)	ニュース記事のクラスタリング結果の評価 (再現率/適合率/F 値 (%))
2009 年 11 月 1 日	1	市長／告示／現職／無所属／再選／投票 (ID:1)，市民 (ID:1,2)	各地の市長選 (ID:1)	100/100/100
	2	市民 (ID:1,2)，追う／藍綬／受章／褒章／黄綬 (ID:2)	秋の褒章 (ID:2)	
2009 年 12 月 19 日	1	削減／気候／首脳／合意／締約／条約／会議／途上／枠組み／国連／先進／コペンハーゲン／排出／交渉／変動／ cop ／途上国／先進国／ガス／来年 (ID:1)	COP15 協議 (ID:1)	100/100/100
2010 年 1 月 13 日	1	昨年 (ID:1,2,3,4)，準備／中国 (ID:1,4)，前日 (ID:2,3,4)，預金 (ID:2,4)，経営 (ID:3,4)，引き上げ／預金準備率／反落／中国人民銀行 (ID:4)	経済 (ID:4)	90.1/91.2/90.7
	2	昨年 (ID:1,2,3,4)，陸山会／東京地検特捜部 (ID:2)，小沢一郎／小沢／一郎 (ID:2,3)，前日 (ID:2,3,4)，預金 (ID:2,4)，日本航空／日航／経営 (ID:3,4)	小沢氏違法献金事件 (ID:2) ／日本航空破綻と再建 (ID:3) ／経済 (ID:4)	
	3	昨年 (ID:1,2,3,4)，前日 (ID:2,3,4)，日本航空／日航／経営 (ID:3,4)	日航経営破綻と再建 (ID:3) ／経済 (ID:4)	
	4	ハイチ／地震 (ID:1)，昨年 (ID:1,2,3,4)，中国／準備 (ID:1,4)	ハイチ地震 (ID:1)	
2010 年 3 月 1 日	1	金メダル／冬季五輪／団体追い抜き／メダル／会式／スケート／閉会／冬季／追い抜く／ 3 月 1 日／バンクーバー／スピード／五輪／銀メダル／スピードスケート (ID:1)	バンクーバー五輪 (ID:1)	100/76.5/86.7
	2	大地震／チリ／津波／南米／地震 (ID:2,3)	チリ大地震 (ID:2) ／国内の津波の影響など (ID:3)	
2010 年 4 月 12 日	1	バンコク／反政府／治安部隊／タイ／デモ／衝突 (ID:1,4,5)，　博之 (ID:3,4,5)，カメラマン／村本／ロイター通信 (ID:4,5)	バンコク騒乱 (ID:4) ／バンコク騒乱で邦人男性死亡 (ID:5)	92.4/90.0/91.2
	2	週末 (ID:1)，バンコク／反政府／治安部隊／タイ／デモ／衝突 (ID:1,4,5)，博之 (ID:3,4,5)	経済 (ID:1)，バンコク騒乱 (ID:4)	
	3	連合 (ID:1,3)，投票 (ID:2,3)，平沼赳夫／平沼／たちあがれる (ID:3)，博之 (ID:3,4,5)	たちあがれ日本が結成 (ID:3)	
	4	ユーロ圏／週末／ギリシャ(ID:1)，連合 (ID:1,3)	経済 (ID:1)	
	5	市長／投票率 (ID:2)，投票 (ID:2,3)	各地の市長選 (ID:2)	
2010 年 5 月 1 日	1	各地 (ID:1,2,4)，水俣市／水俣／水俣病／救済／慰霊／患者 (ID:2)，メーデー (ID:2,3,4)，防止／熊本県／熊本／打つ (ID:2,5)	水俣病犠牲者慰霊式 (ID:2)	97.4/100/98.7
	2	渋滞／行楽／大型連休／連休 (ID:1)，各地 (ID:1,2,4)，上海 (ID:1,3)，キロ (ID:1,4)	ゴールデンウィーク (ID:1)	
	3	上海 (ID:1,3)，メーデー (ID:2,3,4)，上海万博／万博 (ID:3)	上海万博 (ID:3)	
	4	各地 (ID:1,2,4)，キロ (ID:1,4)，メーデー (ID:2,3,4)	メーデー (ID:4)	
	5	熊本県／熊本 (ID:2,5)	普天間問題 (ID:5)	
	6	防止／打つ (ID:2,5)	普天間問題 (ID:5)	